# Peer-Level Calibration of Performance Evaluation Ratings:

# Are There Winners or Losers?

**Peer-Level Calibration of Performance Evaluation Ratings:**

**Are There Winners or Losers?**


## ABSTRACT

In this study we examine the common practice of employee performance rating calibration, the process in which calibration committee members discuss, compare, and potentially adjust supervisors' preliminary subjective employee performance ratings. We highlight the inherent incentive conflict related to calibration between the organization and supervisors, where the organization wants calibration to increase consistency in performance ratings while supervisors are also interested in adjustments that benefit themselves. We show that in peer-level calibration, where supervisors are involved in the calibration of their own employees' ratings, supervisors strategically use this opportunity to influence the calibration process. Specifically, we show that incentive-driven supervisor rating behavior predicts the winners and losers of the peer-level calibration process. The adjustments (or lack thereof) made during the calibration process are not solely driven by the organizational objective of increased rating consistency, but also by supervisors' incentives. Our research has important implications for the designers of performance evaluation and compensation plans. It highlights the importance of the structural design and the composition of calibration committees, and cautions against overestimating the accuracy of post-calibration performance ratings when using them for important decisions such as promotions and resource allocation.



**Keywords:** Calibration; Calibration Committee, Subjectivity; Performance Evaluation, Consistency, Incentive Conflict.

## *1. Introduction*

Performance rating calibration, the process where calibration committee members discuss, compare, and potentially adjust supervisors' preliminary subjective employee performance ratings, has become part of the employee performance evaluation process in many organizations (Lawler, Benson, and McDermott [2012], Hastings [2011]). Mercer's Global Performance Management Survey [2013] shows that of the 1,000 organizations surveyed, more than half (56 percent) use calibration. In this study we examine how supervisor incentives influence the outcomes of the calibration process; i.e., post-calibration employee performance ratings. We investigate the inherent incentive conflict concerning calibration that exists between supervisors and the organization. Organizations want calibration committees to make adjustments that increase consistency in performance ratings (Fox [2009], Sammer [2008]); supervisors, on the other hand, are also interested in adjustments that benefit themselves (Traub [2013]). Understanding that supervisors' incentives in calibration committees are at odds with the organizational objectives and the impact that this has on supervisors' rating choices is important because it triggers organizations to assess and potentially change the use and/or structural design of calibration committees. This can lead to better performance evaluation and compensation processes, which significantly impact organizational success (Baker, Gibbons, and Murphy [1994], Lawler et al. [2012], Kampköter and Sliwka [2015]).

Calibration committees can be structured in different ways. In this study we focus on *peer-level* calibration committees, where peer supervisors meet to calibrate the subjective employee performance ratings of all the supervisors in the calibration committee, including their own (Lawler et al. [2012], Hastings [2011]). Other examples of calibration committee types are *higher-level* calibration committees, where higher-level managers (e.g., segment heads) calibrate

the lower-level supervisors' employee performance ratings (Demeré, Sedatole, and Woods [2018]), and *mixed-level* calibration committees, where both peer supervisors and higher-level managers are part of the calibration committee (Grabner, Künneke, and Moers [2018]). The fact that members of the calibration committee are calibrating their own evaluations of their employees and the evaluations of their direct peers is an important structural design feature of the peer-level calibration process. It provides supervisors with the opportunity to strategically influence the calibration process, which is critical because of the aforementioned incentive conflict. Supervisors have incentives to base their employee performance ratings not purely on employee performance, but to also consider their own personal costs and benefits related to employee performance ratings (Bol [2011]). Specifically, supervisors tend to prefer relatively higher performance ratings (i.e., ratings a little higher than the employees' 'true' performance ratings), especially for the weaker employees, because higher ratings will likely lead to fewer confrontations (Napier and Latham [1986]), greater employee appreciation (Spence and Keeping [2011]), and increased employee motivation (Bol [2011]).

Organizations, on the other hand, want ratings that clearly distinguish among bottom, average, and top performers (Kampkötter and Sliwka [2015], Murphy [1992], Bretz Jr., Milkovich, and Read [1992]). They do not want different performance levels to receive the same performance ratings nor do they want similar performance levels to receive different performance ratings. Both situations would increase perceived unfairness, which is detrimental to employee motivation (Colquitt and Chertkoff [2002], Erdogan [2002]). Rating inconsistencies also negatively influence personnel decisions like promotions and task assignments (Moers [2005]). Thus, organizations want consistency in performance ratings by increasing the consistency of the application of rating criteria and by decreasing rater inaccuracies (Caruso

[2013]). Organizations frequently push down this organizational objective by mandating a set, normal distribution for the overall post-calibration employee performance rating (Cappelli and Tavis [2018], Hastings [2011]).

Considering that organizations want overall performance rating distributions with a non-trivial proportion of the employees rated below the average of the rating scale, and supervisors want to maintain ratings that are relatively higher with few employees rated below the average of the rating scale, the calibration process, and specifically the mandate to meet the organization's required distribution, puts pressure on the calibration committee to make downward adjustments. This, however, does not mean that every performance rating of each supervisor needs to be adjusted downwards. And, as a result, supervisors in peer-level calibration will have incentives to be strategic by suggesting downward adjustments to other supervisors' employee performance ratings while obtaining the preferred higher ratings for their own employees. Accordingly, the calibration process will likely resemble a negotiation where there will be winners, those supervisors able to obtain higher than average post-calibration ratings for their own employees, and losers, those supervisors who leave the calibration process with lower than average ratings for their employees. In this study, we investigate the post-calibration ratings, the adjustments, and the pre-calibration ratings to examine whether strategic incentive-driven behavior by supervisors predicts the winners and losers of the peer-level calibration process.

We hypothesize that supervisors with more political influence will be able to obtain higher post-calibration employee performance ratings. They will use their political influence to convince the rest of the calibration committee members that other supervisors' employee performance ratings should be adjusted downwards while protecting their own employee performance ratings. We also predict that supervisors who will benefit more from higher ratings

will be able to obtain higher post-calibration ratings because they are willing to 'fight harder' (e.g., collect more supporting information, influence peers before the meeting, push harder during the calibration process) to obtain these ratings. Specifically, we hypothesize that supervisors with stronger reputational concerns want to signal strong leadership skills by presenting a well-functioning and, thus, highly rated department (cf. Longenecker, Sims, and Gioia [1987]), and will invest extra time and effort into the calibration process resulting in higher post-calibration employee performance ratings.

Because calibration committees tend to have multiple members (Hastings [2011], Fox [2009]), relationships within the calibration committees likely play an important role in shaping outcomes. We draw from literature on group negotiations showing the importance of formal and informal alliances (Polzer, Mannix, and Neale [1998], Jehn and Bezrukova [2010], Brion and Anderson [2013]). We predict that supervisors will support peer supervisors with whom they have an informal alliance at the expense of supervisors with whom they do not through, for example, not questioning their higher ratings and providing only positive additional performance information on their employees. The presence or absence of peer support matters because of the pressure to reduce ratings to meet the organization's required rating distribution. We predict that supervisors who lack peer support will have a harder time defending their post-calibration ratings during the calibration process, which makes downward adjustments to their ratings more likely. As a result, we hypothesize that supervisors who lack peer support will receive downward adjustments and will have lower post-calibration employee performance ratings.

Finally, considering that confrontation costs are a determinant of relatively higher performance ratings (i.e., leniency bias) in organizations without calibration (Bol [2011]), we examine the influence of supervisors' aversion to confrontation in the peer-level calibration

process. Supervisors who have a higher aversion to confrontation will benefit more from higher post-calibration ratings, as they will reduce the likelihood of confrontations with employees, which are costly to these supervisors. However, supervisors with a higher aversion to confrontations are less willing and able to stand up for their employee performance ratings and confront peer supervisors during the calibration process, again because this is costly to them. As a result, we hypothesize that supervisors with higher aversion to confrontation will enter the calibration process with higher ratings, but will also suffer more downward adjustment due to their unwillingness to confront peers during the calibration process.

We examine our predictions by studying an organization that has used both peer-level calibration committees and a required performance rating distribution since 2011. We were provided access to their 2014 performance evaluation data which contains subjective performance evaluation data for 737 employees submitted by 114 supervisors. Notably, the data includes both pre- and post-calibration performance ratings, information on the twenty-eight peer-level calibration meetings, and demographic data for both the supervisors and employees. We supplement this archival data with survey data on almost all supervisors who participated in the calibration meetings (96.6 percent response rate). The setting and the data provide us with a unique opportunity to examine who are winners or losers of the peer-level calibration process.

We first confirm that the calibration process improves the conformity of the performance rating distribution with the distribution required by the organization's executives. We then test our hypotheses and find evidence consistent with all of them. Specifically, our results show that supervisors' political influence and concern for reputation are positively associated with higher post-calibration employee ratings. We also show that lack of peer support and aversion to confrontation are associated with downward revisions during the calibration process. Taken

together, these results show that peer-level calibration allows supervisors the opportunity to be strategic in the calibration process and, as a result, the adjustments (or lack thereof) made during the calibration process are not solely driven by the organizational objective of increased rating consistency, but also by supervisors' incentives.

This study makes several contributions to the literature on performance evaluation and compensation contracting. First, it contributes to the nascent literature on calibration committees by highlighting the importance of the structural design and the composition of calibration committees. The peer-level structural design, as seen in our setting, has some benefits. For example, supervisors can easily provide additional performance information, provide additional details on how rating criteria have been applied, and answer questions during the meeting (Risher [2014], Sammer [2008]). However, our study shows that by giving supervisors the opportunity to influence the calibration of their own employee ratings, supervisors will act strategically to secure more favorable performance ratings for their own employees. As a result, supervisors might push for an adjustment or refrain from suggesting an adjustment, not because the adjustment or lack thereof improves the consistency of the ratings, but because it helps supervisors secure ratings that will lower their own personal costs or enhance their own personal benefits related to performance evaluation.

These findings have important implications for performance evaluation system designers. Our study indicates that when adding calibration to the performance evaluation process, the organization needs to carefully consider the inherent incentive conflict between supervisors and organizations and the opportunities the structural design of the calibration committee creates for supervisors to prioritize their own incentives over the objectives of the firm. Specifically, peer-level calibration committees provide supervisors with the opportunity to be strategic during the

calibration process and, as a result, organizational objectives related to calibration may become of second importance to the supervisors. Organizations need to weigh the advantages of peer-level calibration against this negative effect.

Our study also highlights the importance of the composition of peer-level calibration committees, which is important given organizations' discretion over choosing calibration committee members. We show that this committee design choice is not trivial because calibration outcomes are influenced by within-calibration-committee relationships. Specifically, our results indicate that supervisor peers that share informal alliances support each other, even at the expense of other supervisors not included in the alliance. This finding shows the importance of carefully considering peer-level calibration committee composition, to avoid putting certain supervisors, and their employees, at a disadvantage.

Second, our study makes a strong contribution to the literature on performance rating accuracy. The majority of papers in this area have focused on identifying inaccuracies and biases (Moers [2005], Bol and Smith [2011]) and its determinants (Bol [2011], Woods [2012]). Only more recently, has the literature focused on how organizations can potentially improve rating accuracy. For example, Grabner et al. [2018] suggest linking the accuracy of supervisors' employee performance ratings to their promotion opportunities, and Bol, Kramer, and Maas [2016] suggest increasing performance information accuracy and performance rating transparency. These approaches might be difficult to implement, therefore, not surprisingly, organizations have been interested in practitioners' claims concerning the "debiasing" effect of calibration committees (e.g., Caruso [2013], Albert [2017]). For example, Risher [2011, p. 275] states about calibration meetings that "The meetings raise the level of honesty, reduce bias and discrimination and provide greater consistency in ratings across an organization". Our study,

however, shows that supervisor incentive-driven bias is still very much present in employee performance ratings even after the use of peer-level calibration committees, thereby casting doubt on practitioners' claims that calibration significantly increases rating consistency.[1]

Our results are relevant for performance evaluation system designers and users because, contrary to the "rosy view" of some professionals (e.g., Risher [2014], Caruso [2013]), our empirical findings suggest that it is important not to overestimate the accuracy of post-calibration committee performance ratings when using them for important decisions such as promotions and resource allocation.

We develop our hypotheses in Section II, and describe our research setting and variables in Section III. Section IV presents the results of our analyses and Section V concludes.

## *2. Background and Hypothesis Development*

Calibration of employee performance ratings is the process in which calibration committee members meet to discuss, compare, justify, and potentially adjust preliminary subjective employee performance ratings (Lawler et al. [2012], Hastings [2011]). The main objective of performance rating calibration is improved performance rating consistency through greater consistency in the application of rating criteria and through decreased rater inaccuracies (Caruso [2013]). Organizations prefer ratings that clearly distinguish among bottom, average, and top performers (Kampkötter and Sliwka [2015], Murphy [1992], Bretz Jr. et al. [1992]). They do not want different performance levels to receive the same performance ratings nor do they want similar performance levels to receive different performance ratings. This objective often results in the organization providing the calibration committees with a mandate to meet a

[1] Unfortunately, we cannot empirically establish whether ratings are more or less consistent/accurate when organizations use calibration committees versus when they do not. For a more detailed discussion see Section V.

predetermined, normal distribution for the post-calibration employee performance rating (Cappelli and Tavis [2018], Hastings [2011]).

Calibration of employee performance ratings has been a common practice for at least the last 10 years (Lawler et al. [2012], Risher [2014], Albert [2017]). For example, the Global Performance Management Survey [2013] found that of the 1,000 organizations from 53 countries surveyed, more than half (56 percent) use calibration meetings. The 2016 WorldatWork survey showed not only that 69 percent of the surveyed organizations use a formal calibration process, but also that 40 (18) percent used calibration for over 5 (10) years (Ledford Jr., Benson, and Lawler [2016]).[2] In a field study examining subjectivity, Lillis, Malina, and Mundy [2018] find that three of the four participating organizations use formal calibration committees.[3]

Despite the widespread use of performance ratings calibration in practice, there is limited academic research on calibration. Further, the studies that do exist focus on calibration committees with different structural designs. For example, Demeré et al. [2018] examine calibration committees where higher-level managers (e.g., segment heads) calibrate the lower-level supervisors' employee performance ratings. They show that calibration leads to a less lenient but more compressed performance rating distribution. Higher-level calibration differs from peer-level calibration because supervisors are not calibrating their own employee ratings, and are therefore not allowed the opportunity to strategically influence the adjustments to their own employee ratings. Grabner et al. [2018] examine mixed-level committees where both peer

[2] Note that only companies that use "ongoing feedback," "rating-less reviews," and "crowd-sourced feedback" are included in this study, which suggests that all of these companies are early adopters of new performance management systems. Hence the percentage of calibration users might be higher than in a randomly selected group of organizations.

[3] Other examples of the prevalence of calibration include: the 2010 Sibson Consulting's worldwide survey on the state of performance management (29 percent of the surveyed organizations used calibration meetings), the 2011 poll of the Society for Human Resource Management (54 percent of the surveyed organizations used calibration meetings), and the 2013 Towers Watson study of performance management practices in the United Kingdom (35 percent of the surveyed organizations used calibration meetings).

supervisors and higher-level managers are part of the calibration committee. They show that the higher-level managers incorporate lower-level supervisors' evaluation behavior in their performance assessments of the lower-level supervisors. This provides lower-level supervisors with incentives to impress the higher-level managers through their employee rating choices. As a result, the structural design of the mixed-level calibration committee creates supervisor incentives (i.e., impressing higher-level managers) that are not present in peer-level calibration committees.

Another structural design feature that differs in our setting compared to these prior studies is the presence of a required employee performance rating distribution. The setting of Demeré et al. [2018] does not mention a required employee rating distribution and the setting of Grabner et al. [2018] has a suggested employee rating distribution that is only a "guideline" that "is not enforced." (Page 22) This is important because the enforcement of a distribution puts more pressure on the calibration committee as a whole to make downward adjustments and consequently increases the need for individual supervisors to be strategic in order to secure higher ratings for their own employees. Thus, we complement prior work on calibration by examining the incentive conflict between the organization and the supervisors in peer-level calibration committees, something that has not been addressed in earlier work, arguably because of the less pronounced role of the incentive conflict in these studies' settings.

## 2.1 SUPERVISOR INCENTIVES IN PERFORMANCE ASSESSMENT

Research in management accounting shows the importance of supervisor incentives in subjective performance evaluation (Grund and Przemeck [2012]). This research shows that subjective performance ratings are not purely based on employee performance but also

influenced by supervisors' own incentives to reduce their personal costs and enhance their personal benefits (Bol [2011], Du, Tang, and Young [2012]). Supervisors are able to consider their own incentives in subjective performance evaluations because outsiders have no way of showing that an individual rating is not the supervisor's 'true' rating.[4,5]

Prior research shows that supervisors, in general, prefer to give relatively higher performance ratings (ratings higher than the 'true' rating), especially to poor performers, because providing employees with higher ratings will likely result in pleasant conversations and appreciation (Bol et al. [2016]). Lower ratings, on the other hand, will likely result in painful conversations, confrontations, and damaged personal relationships (Napier and Latham [1986]). Moreover, lenient performance ratings also tend to align with employees' overestimated self-assessments, which increases fairness perceptions and employee motivation (Bol [2011]).[6,7] These are general tendencies; supervisors also have incentives to provide some employees with lower ratings. For example, the supervisor might want to punish an employee for non-obedient behavior (Longenecker et al. [1987]), build a case for dismissal (Poon [2004]), or reinforce feedback when the employee is not responsive (Taylor, Tracy, Renard, Harrison, and Carroll [1995]). Notwithstanding these exceptions, the collective empirical evidence shows supervisors' preference to provide their employees with higher, more lenient performance ratings in subjective performance assessments (e.g., Moers [2005], Bol [2011]).

[4] This also means that researchers cannot precisely measure the inaccuracies in individual performance ratings. Research can, however, provide theory-consistent empirical evidence on incentive-driven rating biases by showing statistical patterns consistent with rater incentives (see e.g. Bol [2011]).

[5] Another factor that influences subjective employee performance assessments is cognitive biases (e.g., Bol and Smith [2011], Lipe and Salterio [2000]). Cognitive biases are however not incentive driven and are therefore outside the scope of this study.

[6] Psychology research shows that individuals have a tendency to overestimate their abilities relative to their supervisors (McFarlane and Thornton [1986], Harris and Schaubroeck [1988]).

[7] Note that in order for ratings to increase motivation they need to be in line with the employee's self-assessment, not significantly higher. When the ratings are higher than the self-assessed performance suggests, the system will no longer motivate employees to work hard as the link between pay and performance is missing (see Bol [2011]).

As a result, organizations using required performance rating distribution that limit the frequency of high ratings will experience misalignment between supervisors' rating preferences and the organization's required rating distribution. The organizational mandate to comply with the required rating distribution will therefore put pressure on the calibration committees to make predominantly downward adjustments. This, however, does not mean that *each* rating or some ratings of *each* supervisor need to be adjusted downward. As a result, supervisors in peer-level calibration will have incentives to strategically suggest downward adjustments to other supervisors' employee performance ratings while securing the preferred higher ratings for their own employees. Accordingly the calibration process will resemble a negotiation where there will be 'winners', who are supervisors able to obtain higher than average post-calibration ratings for their own employees, and 'losers', supervisors who leave the calibration process with lower than average ratings for their employees. In this study, we investigate the post-calibration ratings, the adjustments, and the pre-calibration ratings of peer-level calibration committees to examine whether strategic incentive-driven behavior by supervisors predicts the winners and losers of the peer-level calibration process.

## 2.2 POLITICAL INFLUENCE

Consistent with the management literature on negotiation (e.g., Kim, Pinkley, and Fragale [2005], Zartman and Rubin [2002], Magee, Galinsky, and Gruenfeld [2007]), we argue that supervisors with more political influence, for example those connected to top management through personal relationships and/or those with more skill for acquiring political capital, will be more successful in obtaining higher employee performance ratings in the calibration process. We argue that supervisors with more political influence, will be more successful in convincing others

of their points of view than other supervisors because going against these supervisors may have negative political repercussions. This is consistent with the budgeting literature where Fisher, Frederickson, and Peffer [2000] find that more powerful negotiators make fewer concessions than less powerful negotiators and therefore end up with more desirable budgets. Given supervisors' preference for higher ratings, we hypothesize that supervisors with more political influence will be more successful in securing higher performance ratings for their own employees. Formally stated:

> *H1: Supervisors' political influence will positively affect the supervisors' post-calibration employee performance ratings.*

## 2.3 REPUTATIONAL CONCERNS

Another reason for supervisors to prefer higher employee performance ratings is because of the signaling value of ratings (Rosaz and Villeval [2012]). That is, supervisors who want to signal strong leadership skills will want to present a well-functioning and highly rated department, independent of whether that is an accurate reflection of true performance. This is consistent with Longenecker et al. [1987] who interviewed 60 executives and found that when individuals outside the department reviewed the ratings, supervisors inflated their ratings to avoid "hanging dirty laundry out in public." (p. 188).

Due to this signaling value, we predict that supervisors who are more actively trying to build or maintain their reputation will invest more time and effort into convincing their peer supervisors that their employees deserve higher ratings. That is, because they will benefit more from higher ratings, they are more willing to 'fight hard' to obtain higher ratings than other supervisors, for example by collecting more supporting information, influencing peers before the meeting, and negotiating more aggressively during the calibration process. Our prediction is consistent with the field interviews in Lillis et al. [2018], in which one of their interview

participants acknowledged that “fighting harder” could result in significantly better outcomes.

Therefore, we hypothesize that supervisors who are more concerned about their reputation will be more willing to invest time and effort into securing higher ratings. Formally stated:

> *H2: Supervisors’ reputational concerns will positively affect the supervisors’ post-calibration employee performance ratings.*

## 2.4 LACK OF PEER SUPPORT

Considering that calibration committees tend to have multiple supervisors (Hastings [2011], Fox [2009]), the role of relationships within the calibration committees needs to be considered. The literature on group negotiation processes shows the importance of formal and informal alliances between different group members (Polzer et al. [1998], Jehn and Bezrukova [2010], Brion and Anderson [2013]). Consistent with this literature we predict that supervisors will build informal alliances through friendships or close working relationships with other calibration committee members and that they will support each other at the expense of those supervisors who are outside the informal alliance. Because of the pressure to reduce employee performance ratings during the calibration process, supervisors without peer support (i.e., no informal alliance) will not have other supervisors endorse their ratings and advocate for their interest, which will likely result in downward rating adjustments. Therefore, we hypothesize that lack of support from peer supervisors will lead to performance rating adjustments and lower post-calibration employee performance ratings. Those supervisors who lack peer support because they are not part of any informal alliance will be the ‘losers’ of the calibration process. Formally stated:

> *H3a: Lack of peer support will positively affect the likelihood that supervisors will receive downward adjustments to their pre-calibration employee performance ratings during the calibration process.*

*H3b: Lack of peer support will negatively affect supervisors' post-calibration employee performance ratings.*

## 2.5 AVERSION TO CONFRONTATIONS

Individuals tend to avoid confrontation because it can trigger negative physical and emotional reactions (e.g., increased stress levels) (Czopp, Monteith, and Mark [2006], Friedman, Tidd, Currall, and Tsai [2000]) and destroy personal relationships (Murphy and Cleveland [1995], Longenecker et al. [1987]). Prior literature shows that supervisors who face higher confrontation costs provide more lenient and compressed employee performance ratings in a setting without calibration (Bol [2011]). We expect these tendencies to persist in a calibration setting. As a result, we predict that supervisors with higher aversion to confrontation will strategically try to avoid personal cost by entering the calibration process with higher ratings.

However, we also expect that the aversion to confrontation will influence the supervisors during the calibration process. We predict that those supervisors with higher aversion to confrontation will be unable to sustain their higher ratings throughout the calibration process. In order to maintain higher employee performance ratings, supervisors need to convince peer supervisors not to make any downward adjustments to their performance ratings. It is unlikely that supervisors with higher aversion to confrontation will be successful in 'standing up' to their peer supervisors like that. Instead, they are more likely to quickly give in when peer supervisors suggest changing their ratings because they want to avoid confrontation. Therefore, we hypothesize that supervisors with stronger aversion to confrontation will enter the calibration process with higher ratings than other supervisors but that they are more likely to receive downward adjustments to their employee ratings. Formally stated:

*H4a: Supervisors' aversion to confrontation will positively affect supervisors' pre-calibration employee performance ratings.*

*H4b: Supervisors' aversion to confrontation will positively affect the likelihood that supervisors will receive downward adjustments to their pre-calibration employee performance ratings during the calibration process.*

## *3. Research Setting and Variables*

### 3.1 THE RESEARCH SITE

We test our hypotheses using performance data from a large publicly-owned Brazilian company. For 2016, the company had total assets of over R$5.7 billion (about $1.71 billion US), net revenues of over R$5.3 billion (about $1.58 billion US), and is considered the Brazilian leader in its industry. The company employs over 16,000 employees located in geographically disperse industrial plants, distribution centers, and administrative offices throughout Brazil. To help standardize performance evaluation within a given hierarchical level across locations, the company implemented a new performance evaluation system in 2010. We focus our study on the managerial level as the evaluation system at this level includes peer-level calibration.[8]

### 3.2 THE PERFORMANCE EVALUATION SYSTEM

The performance evaluation system involves a yearly evaluation of every employee by his or her supervisor using six or seven individual subjective performance measures (i.e., performance results, innovation, initiative, interpersonal skills, communication effectiveness, technical knowledge, and leadership) and one departmental performance measure.[9] Employees receive a rating between one and four for each individual performance measure and the rating on the departmental performance measure is a function of both the department's performance and the employee's individual contribution to the department's performance. Specifically, each

[8] The managerial level includes directors, supervisors, and specialists. Our study focuses on the supervisors' assessments of specialists (hereafter employees).

[9] The decision of which performance measures to use for a particular job function is made at the executive level.

employee within a department receives the same base rating between one and four, contingent on the department's performance, which the supervisor can then adjust one point up or down depending on the employee's contribution to the department's performance. The scores of the individual performance measures are then averaged together and used in conjunction with the individualized departmental performance measure to determine whether an employee is rated as "below expectations", "meets expectations", or "exceeds expectations". As shown in Figure 1, an employee's overall rating is the union of their x-axis and y-axis positions, where the rating on the x-axis is determined by the individualized departmental performance measure and the location on the y-axis is determined by averaged individual performance measures. As participants perform better on both sets of performance measures, they move toward the top right corner of the matrix and their overall performance rating improves.[10] Employees' overall performance ratings influence merit pay, promotion, and job assignment decisions. Additionally, employees rated in the top right cell of Figure 1 are publicly recognized and receive a bonus of R$1,000 (about $300 US) to be used for job-related purchases (e.g., new technology and continuing education courses). To prevent too many employees rated as "exceeds expectations" and not enough rated as "below expectations" and to increase consistency in the application of performance criteria, the firm implemented a required distribution and peer-level calibration meetings, respectively.

The required performance rating distribution is determined by top management. Supervisors have discretion to evaluate their employees as they see fit; however, the overall required company-wide rating distribution is that 10 to 20 percent of the employees receive "exceeds

[10] We find that supervisors attend to the matrix when determining performance ratings. We find that only 5 out of 737 employees (0.7 percent) are given an initial overall rating that does not match what the matrix would suggest given the average of their individual performance rating and the individualized departmental performance rating. Results are unchanged when dropping these five observations.

expectations" ratings, 10 to 20 percent receive "below expectations" ratings and the remaining 60 to 80 percent receive "meets expectations" ratings. The organization implemented this required distribution because the executives wanted the ratings to follow a normal distribution. In particular, the executives wanted to limit the number of employees rated as "exceeds expectations" to better identify true top performers for future promotion and pay increases. Moreover, the executives wanted to ensure that supervisors actually rated their poor performers as "below expectations" to trigger the development of a formal employee improvement plan and to alert the HR department so they could monitor the execution of the improvement plan. The executives of the organization feared that without this performance rating distribution, supervisors would provide too many high ratings and too few low ratings.

Regarding the composition of the peer-level calibration committees, each peer-level calibration meeting consists of at least two supervisors and one or more HR representatives. The HR department, with input from the company executives, assigns each supervisor to a particular meeting. Their goal is to assign supervisors who frequently interact with each other to the same calibration meeting to minimize the level of information asymmetry among supervisors. This often results in supervisors who have a similar job function being assigned to the same calibration meeting. This also means that there is considerable consistency in the calibration committees' composition over time. Although supervisors are occasionally assigned to a different calibration meeting to accommodate work and/or vacation schedules, these instances tend to be the exception rather than the norm.

Each calibration meeting follows a similar set of guidelines. First, the HR representative displays the distribution of the performance ratings of all employees to be discussed during that calibration meeting. Then each supervisor justifies all of his or her employee ratings by

describing how he or she determined the ratings for the individual and individualized departmental performance measures.[11] During this process, the supervisor might be asked to clarify what behavior of the employee led to the ratings, how that behavior is reflective of the specific performance measure, and how often that behavior occurred. After all supervisors in the calibration meeting agree on that employee's evaluation, the ratings are adjusted, if needed. This process is repeated until all supervisors have justified all their employees' ratings.

### 3.3 DATA COLLECTION

We approached the company in 2013 and requested access to their 2014 archival performance evaluation data, which the company provided to us in 2015. This data included both the pre- and post-calibration meeting employee classifications and the pre- and post-calibration meeting scores for the individual and individualized departmental performance measures for over 800 employees provided by 118 supervisors. The company also provided archival calibration meeting data including the list of supervisors that participated in each calibration meeting and the date and length of each calibration meeting. Finally, the company gave us access to demographic data such as gender, age, and tenure for the supervisors and employees.

In addition to collecting archival data, we also surveyed supervisors immediately after their calibration meeting.[12] We sent surveys to the 118 supervisors who participated in the calibration meetings and received 114 responses for a 96.6 percent response rate.[13,14] In the survey,

[11] Supervisors only provide verbal justification for their employee performance ratings.

[12] The survey was conducted in Portuguese. We first created the survey in English and one of the authors and two native Portuguese-speaking Ph.D. students translated the survey into Portuguese. We then pilot-tested the Portuguese version with an employee at the company and made changes based on the feedback. Finally, we shared the Portuguese version with the executives of the organization. They requested that we add, remove, and adjust some of the questions. After several rounds of discussion, we arrived at the final, distributed Portuguese version.

[13] We achieved such a high response rate because top management was interested in receiving feedback on the calibration committee meetings and was therefore willing to send several emails emphasizing the importance of filling out the survey. The HR department also provided us with the date of each calibration committee meeting so

supervisors answered questions about themselves and their perceptions of the other supervisors in their calibration meeting. To better understand the performance evaluation process and to inform our survey questions, we also conducted on-site interviews. In particular, one member of the research team visited the company twice to interview three executives who were responsible for the design of the performance evaluation system and two supervisors who had participated in calibration committee meetings. These interviews provided insight into the calibration meeting process and highlighted recurring themes that were incorporated into our survey.

### 3.4 DEPENDENT VARIABLES

For our analysis, we focus on employees' overall performance ratings. Our first dependent variable, *PRE RATING*, is coded 3, 2, or 1, respectively, based on whether an employee was rated as "exceeds", "meets", or "below expectations" prior to the calibration meeting. Our second dependent variable, *CHANGE DOWN*, is an indicator variable coded 1 (0) if an employee did (did not) receive a downward adjustment during the calibration meeting. Our third dependent variable, *RATING CHANGE*, is calculated as *POST RATING – PRE RATING* and measures the direction and magnitude of each employee's rating change resulting from the calibration meeting. Our fourth dependent variable, *POST RATING*, is coded 3, 2, or 1, respectively, based on whether an employee was rated as "exceeds", "meets", or "below expectations" after the calibration meeting.

---

we could send the survey immediately after the meeting. In addition to top managements' emails, we sent a reminder once a week to those supervisors who had not completed the survey.

[14] For completeness we perform a non-response test on the four supervisors who did not respond. We find that three of the four non-respondents are female while the gender ratio of our respondents is 77 percent male and 23 percent female. We observe no differences in age, job tenure, or firm tenure. Given the low number of non-respondents, we do not feel the difference in the gender ratio between respondents and non-respondents impairs our ability to generalize the results of the survey to the population of interest.

## 3.5 INDEPENDENT VARIABLES

In this section, we discuss the independent variables we use to test our hypotheses. Section IV provides our results using these independent variables as well as alternative specifications of these independent variables.

To test H1, which predicts a positive association between supervisor political influence and post-calibration ratings, we use a proxy for political influence that captures supervisors' ability to network and their connectedness with top executives. To measure networking, we ask supervisors six networking questions from the political skill inventory scale (Ferris et al., [2005]), which measures each supervisor's ability to influence a large number of people through the size and composition of their network.[15] To measure connectedness with top executives, we ask supervisors to indicate which, if any, vice-presidents they had personally talked to within the past year when they wanted to affect the outcome of an important decision. Then, we create a supervisor-level indicator variable coded 1 if a supervisor selected any one of the three most influential vice-presidents from the list and 0 otherwise.[16] We run a single principal component analysis (PCA) using these seven items. All networking and connectedness items load positively on a single factor (*POLITICAL INFLUENCE*) with an eigenvalue of 3.59 that explains 51 percent of variance.[17]

To test H2, which predicts a positive association between supervisors' reputational concerns and post-calibration ratings, we asked supervisors to indicate their perceptions of their peer supervisors' reputational concerns. Specifically, we asked each supervisor to answer the

[15] The responses to all survey questions involved a 7 point Likert Scale with lower (higher) values indicating strong disagreement (agreement) with a given statement.

[16] There are 6 vice-presidents but from our interviews it became clear that not all of them wield the same level of influence. Our contact within the firm indicated that there was a natural divide between the first three and the second three and we therefore used the first three to create our proxy.

[17] All items have a loading score between 0.26 and 0.46. The first networking item, networking time, had the lowest loading score. After dropping this item, all other factor scores are between 0.36 and 0.46. Section IV reports results using this alternative proxy.

following question about each peer supervisor in the calibration committee: "How important it is to _____ that his/her subordinates are perceived as top performers by others in the company?" We then averaged the responses to get a reputational concern score for each supervisor (*REPUTATION*).[18]

To test H3a (H3b), which predicts supervisors' lack of peer support is positively (negatively) associated with the likelihood of downward adjustments (post-calibration meeting ratings), we use a proxy that captures lack of peer support. We reason that supervisors who work in the same functional area (e.g., accounting, finance, sales) are likely to build informal alliances with each other because of their close working relationship, which will result in support during the calibration process. Supervisors without these close colleagues in their calibration committee will therefore likely lack this support. As a result, we code as 1 those supervisors who are the only one from their functional area in their calibration meeting, contingent upon the calibration meeting having more than two supervisors, 0 otherwise. (*LOW PEER SUPPORT).*

To test H4a (H4b), which predicts supervisors' aversion to confrontation is positively associated with pre-calibration meeting ratings (the likelihood of their ratings being adjusted downward), we asked supervisors to indicate their perceptions of their peer supervisors' confrontation avoidance. Specifically, we asked each supervisor in the calibration committee to answer three questions about their fellow supervisors' confrontation avoidance behaviors. For each supervisor, we averaged the peer supervisors' responses to each of the three questions. We use PCA to combine these three questions and get a single confrontation avoidance score for each supervisor. All three items load positively on a single factor with an eigenvalue of 2.37 that

[18] The final set of survey questions is the result of several rounds of back and forth between the researchers and the company. In the original survey, we asked multiple questions about reputational concerns; however, senior management would not allow us to include those questions in the survey.

explains 79 percent of variance (*AVOID CONFRONTATION*).[19] Table 1 provides a list of all questions used to create our variables.

### 3.6 CONTROL VARIABLES

At the calibration meeting level, we control for the size of the calibration meeting by including a variable measuring the number of supervisors present in each meeting (*CAL SIZE*). At the employee level we control for job type (*EMP JOB*) and the region of Brazil in which the employee is located (*EMP REGION*).[20]

## *4. Results*

### 4.1 DESCRIPTIVE STATISTICS AND CORRELATION MATRIX

Table 2 provides the descriptive statistics for our dependent and independent variables. The average post-calibration rating is significantly lower than the average pre-calibration rating (2.07 vs. 2.13, $p < 0.01$, two-tailed, non-tabulated) and the mean rating change is negative, suggesting most adjustments were downward adjustments. We also note that in 2014, twenty-eight calibration meetings took place.[21] The meetings had between two and nine supervisors, with an average of 4.44 supervisors.

Table 3 provides the correlation of our variables. We observe that *PRE RATING* is negatively correlated with *RATING CHANGE* (corr. coef. = -0.36; $p < 0.01$). This suggests that employees with higher (lower) initial ratings are more likely to receive downward (upward) adjustments.

[19] Although this hypothesis focuses on confrontation avoidance and not perceived confrontation avoidance we used perceptions of peers as we were afraid that supervisors might not truthfully answer these questions about themselves.

[20] The company identified six major job types, which are used to create the indicator variable *EMP JOB*. The company divides its operations into four different regions, which are used to create the indicator variable *EMP REGION*.

[21] For one calibration meeting only one supervisor provided survey data. We exclude this observation from our analysis because some of our independent variables involve the perception of other supervisors in the calibration meeting. As such, our final analyses include 27 calibration-meeting observations.

We note *PRE RATING* is positively correlated with *POST RATING* (corr. coef. = 0.84; $p < 0.01$), which suggests that employees with higher (lower) initial ratings are more likely to leave the calibration with higher (lower) ratings. We also note that *REPUTATION* and *AVOID CONFRONTATION* are positively correlated with *PRE RATING* (corr. coef. = 0.19 and 0.24; $p = 0.04$ and 0.01, respectively) while *REPUTATION* (*LOW PEER SUPPORT*) is positively (negatively) correlated with *POST RATING* (corr. coef. = 0.19 and -0.19; $p = 0.04$ and 0.04, respectively). This pattern of correlations is consistent with H2, H3, and H4.

## 4.2 RATING DISTRIBUTIONS

To confirm that the employee performance ratings distribution after the calibration meets the required distribution we examine the distribution of the pre- and post-calibration performance ratings. Panel A of Table 4 indicates the majority of employees (68.9 percent) were rated "meets expectations" prior to the calibration meeting. Further, more employees were rated "exceeds expectations" and fewer employees were rated "below expectations" than the required distribution suggests (22.0 percent and 9.1 percent, respectively, compared to 10-20 percent). Panel B reveals that there were 80 ratings changes; that is, 10.9 percent of employees received a rating adjustment during calibration. Panel B also reveals that most of the rating changes (77.5 percent) were downward. The majority of the downward rating changes (71 percent) reclassified employees from "exceeds" to "meets expectations". The majority of upward revisions (61 percent) reclassified employees from "meets" to "exceeds expectations". As expected, the rating distribution post-calibration is better aligned with the distribution required by the organization. Panel A of Table 4 shows that after calibration, 10.6 (72.2) [17.2] percent of employees are rated as "below" ("meets") ["exceeds expectations"], consistent with the required distribution of 10-20

(60-80) [10-20] percent of employees rated as "below" ("meets") ["exceeds expectations"]. Panel C provides detail of the ratings at the calibration level meeting. We observe that 20 (3) [5] of the calibration meetings had, on average, downward (upward) [no] rating changes.

## 4.3 MODEL SPECIFICATION

We utilize multiple regression models to test our hypotheses. Our general model is:

$$Rating = \alpha_0 + \alpha_1(POLITICAL\ INFLUENCE) + \alpha_2(REPUTATION) + \alpha_3(LOW\ PEER\ SUPPORT) + \alpha_4(AVOID\ CONFRONTATION) + \alpha_5(CAL\ SIZE) + \alpha_{6-10}(EMP\ JOB) + \alpha_{11-13}(EMP\ REGION) + \epsilon$$

where *Rating* is either *PRE RATING*, *CHANGE DOWN*, *RATING CHANGE*, or *POST RATING*. *Rating*, *EMP JOB*, and *EMP REGION* are at the employee level, *CAL SIZE* is at the calibration meeting level, and all other variables are at the supervisor level. We cluster standard errors by calibration meeting for all regressions.

To examine pre-calibration ratings, we utilize an ordered probit with *PRE RATING* as the dependent variable because the variable is ordered with 1 being the lowest rating and 3 being the highest rating (Model 1). To examine rating adjustments, we run three regressions. The first and second are logit regressions with *CHANGE DOWN* as the dependent variable without and with a control for *PRE RATING*, given the significant negative correlation between *RATING CHANGE* and *PRE RATING* (Models 2 and 3, respectively). The third regression is an ordinary least squares (OLS) regression with *RATING CHANGE* as the dependent variable that also controls for *PRE RATING* (Model 4). Finally, we run an ordered probit with *POST RATING* as the dependent variable without and with a control for department performance (Models 5 and 6, respectively). Using multiple models allows us to examine the winners and losers of the calibration process (i.e., supervisors who left the calibration meetings with higher/lower average

employee performance ratings, respectively). Our approach also allows us to examine how final ratings were achieved (i.e., whether the winners (losers) entered the calibration process with higher (lower) employee performance ratings and/or received rating changes during calibration).

## 4.4 HYPOTHESIS TESTING

Table 5 presents results of our hypothesis tests. H1 predicts that supervisors' political influence is positively associated with post-calibration employee performance ratings. Model 5 reveals support for this hypothesis in the form of a positive association between *POST RATING* and *POLITICAL INFLUENCE* (coef. = 0.10; $p < 0.01$, one-tailed). Model 1 shows that these supervisors enter the calibration meetings with higher pre-calibration scores (coef. = 0.09; $p = 0.03$, two-tailed) and Model 2 indicates that because of their political influence, these supervisors do not receive significant downward adjustments (coef. = 0.00; $p = 0.97$, two-tailed). In addition, we compare the regression coefficient of *POLITICAL INFLUENCE* across Model 1 (*PRE RATING*) and Model 5 (*POST RATING*) and find that the difference between the coefficients is not significant ($Chi^2 = 0.05$; $p = 0.82$, two-tailed, non-tabulated). We find similar results using alternative specifications of *POLITICAL INFLUENCE*.[22] Taken together, this suggests that supervisors with higher political influence are winners of the calibration process in that they leave the calibration process having secured higher employee performance ratings relative to their peers.

H2 predicts that supervisors' reputational concerns are positively associated with post-calibration employee performance ratings. Model 5 reveals support for this hypothesis in the

[22] We rerun our analyses dropping the first networking item, networking time, due to its relatively low loading in the PCA. We also rerun our analyses using the networking questions as one variable and the connectedness question as another separate variable. Results are robust to these alternative specifications (all $p \leq 0.02$, one-tailed). We also average the responses to the questions together instead of using PCA and find consistent results ($p < 0.01$, one-tailed).

form of a positive association between *POST RATING* and *REPUTATION* (coef. = 0.19; p < 0.01, one-tailed). Model 1 shows that these supervisors enter the calibration meetings with higher pre-calibration scores (coef. = 0.21; p < 0.01, one-tailed) and Model 2 indicates that because of their reputational concerns, these supervisors do not receive significant downward adjustments (coef. = -0.01; p = 0.98, two-tailed). In addition, we compare the regression coefficient of *REPUTATION* across Model 1 (*PRE RATING*) and Model 5 (*POST RATING*) and find that the difference between the coefficients is not significant ($Chi^2$ = 0.26; p = 0.61, two-tailed, non-tabulated). This suggests that supervisors with higher reputational concerns are winners of the calibration process in that they leave the calibration process having secured higher employee performance ratings relative to their peers.

H3a predicts that supervisors' lack of peer support will increase the likelihood of receiving downward adjustments. Evidence consistent with this hypothesis would be a positive association between *LOW PEER SUPPORT* and *CHANGE DOWN* in Model 2. We find evidence consistent with this as indicated by the significant positive coefficient for *LOW PEER SUPPORT* (coef. = 0.85; p = 0.02, one-tailed). We also find consistent evidence in Models 3 and 4. In particular, *LOW PEER SUPPORT* is positively associated with *CHANGE DOWN* and negatively associated with *RATING CHANGE* controlling for *PRE RATING* (coef. = 1.31 and -0.06; p < 0.01 and = 0.03, one-tailed, respectively), again suggesting that supervisors without peer supervisor support receive downward adjustments. We also hypothesize that lack of peer support affects post-calibration ratings (H3b). The negative association between *LOW PEER SUPPORT* and *POST RATINGS*, which we find in Model 5 (coef = -0.24, p = 0.01, one-tailed), supports this hypothesis. We note that *LOW PEER SUPPORT* is not associated with *PRE RATING* in Model 1, suggesting that these supervisors do not have significantly different employee performance

ratings (compared to their peers' employee performance ratings) when entering the calibration meeting. A comparison of the coefficients on *LOW PEER SUPPORT* across Model 1 (*PRE RATING*) and Model 5 (*POST RATING*) confirms that the adjustments received were significant as the coefficient on *POST RATING* is lower than the coefficient on *PRE RATING* ($Chi^2 = 1.95$; $p = 0.08$, one-tailed, non-tabulated). We find similar results using alternative specifications of *LOW PEER SUPPORT*.[23] Taken together, this suggests that supervisors without peer support are losers of the calibration meetings.

H4a predicts that supervisors' aversion to confrontation is positively associated with the supervisors' pre-calibration meeting ratings. The significant positive coefficient for *AVOID CONFRONTATION* (coef 0.09; $p = 0.02$, one-tailed) in Model 1 supports this hypothesis. H4b predicts supervisors with higher aversion to confrontation receive more downward adjustments than supervisors without a strong aversion to confrontation. To investigate, we examine the association between *AVOID CONFRONTATION* and *CHANGE DOWN* in Model 2. The significant positive coefficient for *AVOID CONFRONTATION* (coef. = 0.30; $p = 0.02$, one-tailed) supports H4b. Models 3 and 4 yield additional support for H4b. In particular, *AVOID CONFRONTATION* is positively associated with *CHANGE DOWN* and negatively associated with *RATING CHANGE* controlling for *PRE RATING* (coef. = 0.17 and -0.02; $p = 0.10$ and $= 0.04$, one-tailed, respectively). A comparison of the coefficients on *AVOID CONFRONTATION* across Model 1 (Pre Rating) and Model 5 (Post Rating) confirms the adjustments received were significant ($Chi^2 = 8.06$; $p < 0.01$, one-tailed, non-tabulated). We find similar results using an

[23] We test the robustness of H3 using two alternative specifications of *LOW PEER SUPPORT*. For the first alternative specification we code supervisors who are the only ones in the company working in a particular functional area as one. For the second alternative specification, we code those supervisors who are from a small functional area (fewer than 4 supervisors) as one. Both H3a and H3b are supported with these alternative specifications (all $p < 0.02$, one-tailed using either specification).

alternative specification of *AVOID CONFRONTATION*.[24] Taken together, this suggests that supervisors with high levels of confrontation avoidance are losers of the calibration process.

## 4.5 ADDITIONAL EVIDENCE AND ROBUSTNESS

*4.5.1 Objective Departmental Ratings.* One alternative explanation for our findings is that higher (lower) post-calibration employee performance ratings are not driven by supervisors' incentive-driven rating behavior but by real differences in employee performance. For example, the higher ratings for supervisors with high political influence could arise if these influential supervisors have better-performing employees. To test this alternative explanation, we rerun our analysis with *POST RATING* and include the departmental objective performance measure as a control variable. This objective performance measure indicates the average employee performance level of the department and therefore is not influenced by supervisors' incentive-driven rating behavior. If differential employee performance explains our findings, then we expect our results to be weakened by the inclusion of this control variable. Model 6 in Table 5 provides the results. We first note that the departmental objective performance measure is positively associated with employees' overall ratings, as expected (coef. = 0.45; $p < 0.01$, two-tailed). However, we still find that *POLITICAL INFLUENCE* and *REPUTATION* are both significantly positively associated with *POST RATING* (coef = 0.08 and 0.12; $p < 0.01$ and = 0.02, one-tailed, respectively). Moreover, *LOW PEER SUPPORT* continues to be negatively associated with *POST RATING* (coef = -0.22; $p = 0.01$, one-tailed). These results suggest that the higher (lower) post-calibration ratings for employees of supervisors with higher political influence and reputational concerns (lack of peer support) are not solely driven by higher (lower)

[24] We test the robustness of H4a and H4b using an alternative specification of *AVOID CONFRONTATION*. We average the responses of the three questions together instead of combining them using PCA. Results are robust to this alternative specification (both H4a and H4b are supported with $p < 0.05$, one-tailed).

performing employees and that supervisor incentives influence the calibration process and outcomes.

*4.5.2 Evidence of the Importance of Peer support.* Hypothesis 3a and 3b show that lack of peer support leads to downward adjustments and lower post-calibration ratings. In this additional analysis section, we examine whether having peer support positively influences calibration outcomes. We argue that those supervisors who have at least one other supervisor from their functional area in their calibration meeting are able to build informal alliances that lead to willingness to support each other during the calibration process, which results in higher post-calibration employee performance ratings. To examine this, we again focus on the composition of the calibration committees. Within each calibration meeting, we code supervisors that have at least one other supervisor in their functional area as one (zero otherwise). Panel A of Table 6 reveals through two-tailed t-tests that supervisors with support from peers have directionally higher pre-calibration ratings (2.15 vs. 2.10; t = 1.32; p = 0.19, two-tailed) and, despite having higher pre-ratings, they receive *fewer* downward adjustments (0.06 vs. 0.11; t = 2.48 p = 0.01, two-tailed). Together this results in higher post-calibration ratings (2.10 vs. 2.02; t = 2.21; p = 0.03, two-tailed).

We also examine whether calibration committees with supervisors who all work in the same functional area differ from calibration committees with supervisors from a variety of functional areas. We code those calibration committees that are homogenous in nature (i.e., all supervisors in the committee work in the same functional area) and compare their pre- and post-calibration ratings to those who are not. Panel B of Table 6 reveals that homogenous calibration committees have higher pre-calibration ratings than non-homogenous calibration committees (2.19 vs. 2.09; t

= 2.60; p = 0.01, two-tailed). Despite these higher ratings, they receive *fewer* downward adjustments (0.05 vs. 0.10; t = 2.44; p = 0.02, two-tailed). Together this results in higher post-calibration ratings (2.16 vs. 2.00; t = 4.10; p < 0.01, two-tailed). These analyses provide evidence that close working relationships that result in peer support provide supervisors with an advantage during the calibration process.

*4.5.3 Alternative Model Specification.* We also test the robustness of our results using alternative regression specifications. In particular we use a multilevel mixed-effects model to acknowledge that our independent variables are at three different hierarchical levels. In particular, our multilevel structure consists of 737 employee observations (level 1) nested in 113 supervisor observations (level 2) nested in 27 calibration meeting observations (level 3). The advantage of using this alternative regression specification is that it decreases the risk of Type 1 errors because the standard errors are not underestimated (Raudenbush and Bryk [2002]). Specifically, we use the following model where *i*, *j*, and *k* are the subscripts for employee, supervisor, and calibration meeting, respectively and *DV* is either *PRE RATING*, *CHANGE DOWN*, *RATING CHANGE*, or *POST RATING*.

$$Level\ 1\text{: } DV_{i,j,k} = \alpha_{0,j,k} + \alpha_{1,j,k}(EMP\ JOB) + \alpha_{2,j,k}(EMP\ LOCATION) + e_{i,j,k}$$

$$Level\ 2\text{: } \alpha_{0,j,k} = \beta_{0,0,k} + \beta_{0,1,k}(POLITICAL\ INFLUENCE\ ) + \beta_{0,2,k}(REPUTATION) + \beta_{0,3,k}(LOW\ PEER\ SUPPORT) + \beta_{0,4,k}(AVOID\ CONFLICT) + u_{0,i,j}$$

$$Level\ 3\text{: } \beta_{0,0,k} = \gamma_{0,0,0} + \gamma_{0,0,1}(CAL\ SIZE) + v_{0,0,k}$$

We provide the results of the alternative regression specification in Table 7 and note that our results are largely unchanged. This shows that our results are robust to using multilevel mixed-effects models.

## *5. Summary and Conclusion*

In this paper we examine the role of supervisor incentives in *peer-level* calibration committees. Peer-level calibration committees, where supervisors are involved in the calibration of their own employee performance ratings, provide supervisors opportunities to influence calibration outcomes in their own favor. We show that some supervisors are indeed able to secure more advantageous employee performance ratings. Specifically, our results suggest that supervisors with more political influence and supervisors with greater reputational concerns strategically enter the calibration process with higher ratings and are able to maintain these higher ratings. We also find evidence of incentive-driven downward adjustments during the calibration process. Our results indicate that supervisors who lack peer support and supervisors who are high on confrontation avoidance are more likely to receive downward adjustments.

These findings have important implications for performance evaluation systems designers. Our study indicates that when adding calibration to the performance evaluation process, designers need to be cognizant of the inherent incentive conflict related to calibration between supervisors and the organization. When the organization uses peer-level calibration, supervisors will have incentives to be strategic during the calibration process to secure higher performance ratings. This will likely come at the expense of the organizational objective related to calibration: improving rating consistency. It is therefore important to carefully weigh the organizational costs and benefits of each calibration design structure. The designers of the performance evaluation system also need to carefully consider the composition of the calibration committees. Our results indicate that not considering group dynamics, specifically the extent to which supervisors are able to build informal alliances, can put certain supervisors, and thereby their employees, at a disadvantage.

As with any study, the limitations of our study provide opportunities for future research. Although our results clearly indicate the *presence* of incentive-driven supervisor behavior, which allows us to conclude that several consultants are overselling the extent to which calibration can eliminate rater bias (e.g. Caruso [2013], Albert [2017]), it is still an open question whether introducing calibration into the performance evaluation process increases rating accuracy compared to a performance evaluation process without calibration. Although calibration leads to the required performance rating distribution, which arguably limits how lenient ratings are on average, this does not mean that ratings are more accurate on an individual basis. Arguably, it is more problematic when some ratings are inflated and some are deflated, as documented in this study, compared to a situation where all ratings are lenient, a pattern that has been repeatedly documented in settings without calibration committees (e.g., Moers [2005], Bol [2011]). Future research can develop a sophisticated field experiment to examine whether calibration improves rating accuracy.

Moreover, in this study we have focused on incentive-driven behavior, but subjective performance ratings are also subject to unconscious biases. Supervisors, like everyone else, use cognitive shortcuts based on stereotypes in decision making, which results in more advantageous/disadvantageous behavior towards certain groups of people (Banaji and Greenwald [2016], Uhlmann and Cohen [2005]). Calibration committees can potentially help combat unconscious biases by drawing attention to the possibility that one might unconsciously apply biases, and by creating a more deliberate decision making process (i.e., one less based on cognitive shortcuts). We leave it to future research to examine the effect of calibration on unconscious biases in subjective performance evaluation.

Our study also highlights some interesting opportunities for future research. Our study

suggests that we cannot speak of *the* consequences of calibration; the effect that calibration has on the performance evaluation process is influenced by the structural design of the calibration committees. For example, Demeré et al. [2018], in a setting with higher-level calibration committees, find no evidence of strategic behavior, while we find abundant evidence of strategic behavior in our peer-level calibration setting. Therefore, the extent to which managers engage in incentive-consistent strategic behavior during calibration is likely a function of the type of calibration occurring. More research is warranted to gain a full understanding of how the structural design of the calibration process influences the calibration outcomes.

# REFERENCES


ALBERT, L. 'More Companies Using Calibration to Assess Talent.' Talent Management and HR, June 16, 2017. Available at: https://www.tlnt.com/more-companies-using-calibration-to-assess-talent/

BANAJI, M. R., and A. G. GREENWALD. Blindspot: Hidden Biases of Good People. Bantam, 2016.

BAKER, G., GIBBONS, R., & MURPHY, K. J. 'Subjective Performance Measures in Optimal Incentive Contracts.' The Quarterly Journal of Economics, 109 (1994): 1125-56.

BOL, J. C. 'The Determinants and Performance Effects of Managers' Performance Evaluation Biases.' The Accounting Review 86 (2011): 1549–75.

BOL, J. C., and S. D. SMITH. 'Spillover Effects in Subjective Performance Evaluation: Bias and the Asymmetric Influence of Controllability.' The Accounting Review 86 (2011): 1213-30.

BOL, J. C., S. KRAMER, and V. S. MAAS. 'How Control System Design Affects Performance Evaluation Compression: The Role of Information Accuracy and Outcome Transparency.' Accounting, Organizations and Society 51 (2016): 64–73.

BRETZ Jr, R. D., G. T. MILKOVICH, and W. READ. 'The Current State of Performance Appraisal Research and Practice: Concerns, Directions, and Implications.' Journal of Management 18 (1992): 321-352.

BRION, S., & ANDERSON, C. 'The Loss of Power: How Illusions of Alliance Contribute to Powerholders' Downfall.' Organizational Behavior and Human Decision Processes 121 (2013), 129-39.

CAPPELLI, P., and A. TAVIS. 'HR Goes Agile.' Harvard Business Review, March-April, 2018. Available at: https://hbr.org/2018/03/the-new-rules-of-talent-management

CARUSO, K. N. 'A Practical Guide to Performance Calibration: A Step-by-Step Guide to Increasing The Fairness and Accuracy of Performance Appraisal.' viaPeople, Inc., October, 2013. Available at: http://cdn2.hubspot.net/hub/91252/file-338193964-pdf/Practical_Guide_to_Performance_Calibration_October_2013.pdf

COLQUITT, J. A., and J. M. CHERTKOFF. 'Explaining Injustice: The Interactive Effect of Explanation and Outcome on Fairness Perceptions and Task Motivation.' Journal of Management 28 (2002): 591–610.

CZOPP, A. M., M. J. MONTEITH, and A. Y. MARK. 'Standing up for a Change: Reducing Bias through Interpersonal Confrontation.' Journal of Personality and Social Psychology 90 (2006): 784–803.

DEMERÉ, W., K. L. SEDATOLE, and A. WOODS. 'The Role of Calibration Committees in Subjective Performance Evaluation Systems.' Management Science (2018): In Press.

DU, F., G. TANG, and S. M. YOUNG. 'Influence Activities and Favoritism in Subjective Performance Evaluation: Evidence from Chinese State-Owned Enterprises.' The Accounting Review 87 (2012): 1555–88.

ERDOGAN, B. 'Antecedents and Consequences of Justice Perceptions in Performance Appraisals.' Human Resource Management Review 12 (2002), 555–78.

FERRIS, G. R., D. C. TREADWAY, R. W. KOLODINSKY, W. A. HOCHWARTER, C. J. KACMAR, C. DOUGLAS, and D. D. FRINK. 'Development and Validation of the Political Skill Inventory.' Journal of Management 31 (2005.): 126–52.

FISHER, J. G., J. R. FREDERICKSON, and S. A. PEFFER. 'Budgeting: An Experimental Investigation of the Effects of Negotiation.' The Accounting Review 75 (2000): 93–114.

FOX, A. 'Curing What Ails Performance Reviews.' Society for Human Resource Management (SHRM), January 1, 2009. Available at: https://www.shrm.org/hr-today/news/hr-magazine/pages/0109fox.aspx

FRIEDMAN, R. A., S. T. TIDD, S. C. CURRALL, and J. C. TSAI. 'What Goes around Comes around: The Impact of Personal Conflict Style on Work Conflict and Stress.' International Journal of Conflict Management 11 (2000): 32–55.

GRABNER, I., J. KÜNNEKE, and F. MOERS. 'How to Mitigate Bias in Performance Evaluations: An Analysis of the Consequences of Supervisors' Evaluation Behavior.' Unpublished paper, Maastricht University, 2018. Available at: https://business.illinois.edu/accountancy/wp-content/uploads/sites/12/2016/04/III-Grabner-Kunneke-Moers.pdf

GRUND, C., and J. PRZEMECK. 'Subjective Performance Appraisal and Inequality Aversion.' Applied Economics 44 (2012): 2149–55.

HARRIS, M. M., and J. SCHAUBROECK. 'A Meta-Analysis of Self-supervisor, Self-peer, and Peer-supervisor Ratings.' Personnel Psychology 41 (1988): 43–62.

HASTINGS, R. 'Survey: Most Large Firms Calibrate Performance.' Society for Human Resource Management (SHRM), December 15, 2011. Available at: https://www.shrm.org/ResourcesAndTools/hr-topics/employee-relations/Pages/CalibratePerformance.aspx

JEHN, K. A., & BEZRUKOVA, K. 'The Faultline Activation Process and the Effects of Activated Faultlines on Coalition Formation, Conflict, and Group Outcomes.' Organizational Behavior and Human Decision Processes 112 (2010): 24-42.

KAMPKOTTER, P., and D. SLIWKA. 'The Complementary Use of Experiments and Field Data to Evaluate Management Practices: The Case of Subjective Performance Evaluations.' Journal of Institutional and Theoretical Economics 172 (2015): 354-89.

KIM, P. H., R. L. PINKLEY, and A. R. FRAGALE. 'Power Dynamics in Negotiation.' Academy of Management Review 30 (2005): 799–822.

LAWLER, E., G. BENSON, and M. MCDERMOTT. 'Performance Management and Reward Systems.' WorldatWork Journal 21 (2012):19-28. Available at: https://ceo.usc.edu/files/2016/10/2012-10-G12-10-617-Performance_Management_Reward_Systems.pdf

LEDFORD Jr., G. E., G. BENSON, and E. LAWLER. 'Cutting-edge Performance Management: 244 Organizations Report on Ongoing Feedback, Ratingless Reviews and Crowd-sourced Feedback.' WorldatWork Research, Center for Effective Organizations, August, 2016. Available at: https://www.worldatwork.org/docs/research-and-surveys/research-report-cutting-edge-performance-management.pdf

LILLIS, A. M., M. A. MALINA, and J. MUNDY. 'Rendering Subjectivity Informative in Performance Measurement and Reward Systems: Field Study Insights. Unpublished paper, University of Melbourne, 2018. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2998471

LIPE, M. G., & SALTERIO, S. E. 'The Balanced Scorecard: Judgmental Effects of Common and Unique Performance Measures'. The Accounting Review 75 (2000): 283-98.

LONGENECKER, C. O., H. P. SIMS Jr, and D. A. GIOIA. 'Behind the Mask: The Politics of Employee Appraisal.' The Academy of Management Executive (1987-1989) 1 (1987): 183–93.

MAGEE, J. C., A. D. GALINSKY, and D. H. GRUENFELD, D. H. 'Power, Propensity to Negotiate, and Moving First in Competitive Interactions.' Personality and Social Psychology Bulletin 33 (2007): 200–12.

MCFARLANE S. L. and G. C. THORNTON. 'Effects of Gender on Self- and Supervisory Ratings.' Academy of Management Journal 29 (1986): 115–29.

MERCER. 2013 Global Performance Management Survey Report: Executive Summary. 2013. Available at: https://www.mercer.com/content/dam/mercer/attachments/global/Talent/Assess-BrochurePerfMgmt.pdf

MOERS, F. 'Discretion and Bias in Performance Evaluation: The Impact of Diversity and Subjectivity.' Accounting, Organizations and Society 30 (2005): 67-80.

MURPHY, K. J. 'Performance Measurement and Appraisal: Motivating Managers to Identify and Reward Performance.' In: BRUNS, W. J. (ed.): Performance Measurement, Evaluation, and Incentives. Boston: Harvard Business School Press, 1992, 37-62.

MURPHY, K. R., and J. CLEVELAND. Understanding Performance Appraisal: Social, Organizational, and Goal-based Perspectives. Thousands Oaks: Sage Publications, Inc, 1995.

NAPIER, N. K., and G. P. LATHAM. 'Outcome Expectancies of People Who Conduct Performance Appraisals.' Personnel Psychology 39 (1986): 827–37.

POLZER, J. T., MANNIX, E. A., and NEALE, M. A. 'Interest Alignment and Coalitions in Multiparty Negotiation.' Academy of Management Journal 41 (1998): 42-54.

POON, J. M. 'Effects of Performance Appraisal Politics on Job Satisfaction and Turnover Intention.' Personnel Review 33 (2004): 322-34.

RAUDENBUSH, S. W., and A. S. BRYK. Hierarchical Linear Models: Applications and Data Analysis Methods. Vol. 1. Thousands Oaks: Sage Publications, Inc, 2002.

RISHER, H. 'Reward Management Depends Increasingly on Procedural Justice.' Compensation & Benefits Review 46 (2014): 135–38.

RISHER, H. 'Getting Performance Management on Track.' Compensation & Benefits Review 43 (2011): 273–81.

ROSAZ, J., and M. C. VILLEVAL. 'Lies and Biased Evaluation: A Real-effort Experiment.' Journal of Economic Behavior and Organization 84 (2012): 537-49.

SAMMER, J. 'Calibrating Consistency.' Society for Human Resource Management (SHRM). January 1, 2008. Available at: https://www.shrm.org/hr-today/news/hr-magazine/pages/1hr%20management%20agenda.aspx

SIBSON CONSULTING. 2010 Survey. 2010. Available at: http://www.sibson.com/media/1612/2010spm.pdf

SPENCE, J. R., & KEEPING, L. 'Conscious Rating Distortion in Performance Appraisal: A Review, Commentary, and Proposed Framework for Research.' Human Resource Management Review 21 (2011): 85-95.

TAYLOR, M. S., TRACY, K. B., RENARD, M. K., HARRISON, J. K., & CARROLL, S. J. 'Due Process in Performance Appraisal: A Quasi-Experiment in Procedural Justice.' Administrative Science Quarterly 40 (1995): 495-523.

TOWERS WATSON. 2013 Towers Watson Study of Performance Management Practices. 2013.

TRAUB, L. Bias in Performance Management Review Process: Creating an Inclusive Talent Pipeline by

Understanding Our Filters. A Cook Ross Publication, 2013. Available at: http://www.cookross.com/docs/unconsciousbiasinperformance2013.pdf

UHLMANN, E. L., and G. L. COHEN. 'Constructed Criteria: Redefining Merit to Justify Discrimination.' Psychological Science 16 (2005): 474-80.

WOODS, A. 'Subjective Adjustments to Objective Performance Measures: The Influence of Prior Performance.' Accounting, Organizations and Society 37 (2012): 403-25.

ZARTMAN, I. W., and J. Z. RUBIN. 'The Study of Power and the Practice of Negotiation.' In: ZARTMAN, I. W. and J. Z. RUBIN (eds.): Power and Negotiation. Ann Arbor: University of Michigan Press, 2002, 3–28.

**FIGURE 1**
**Performance Evaluation Matrix**

| Average Individual Performance Measures | Individualized Departmental Performance Measure 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 3.5 – 4.0 | N/A | Meets | Above | Above |
| 2.5 – 3.4 | N/A | Meets | Meets | Meets |
| 1.5 – 2.4 | Below | Below | Below | Below |
| 0.0 – 1.4 | Below | Below | Below | Below |

Middle management employees are evaluated using six or seven individual-level performance measures and one individualized department-level performance measure. Each employee receives a score between one and four on each of the individual-level performance measures. The individual-level performance measures are averaged together and this average determines where an employee falls along the y-axis of the matrix. The individualized department-level performance measure is a function of both the employee's department performance and the employee's individual contribution to the department's results. That is, each employee within a department will receive the same base score between one and four contingent on the department's performance. The department supervisor can then adjust the score of each individual employee up or down one point, based on his or her contribution to the department's performance. This determines where an employee falls along the x-axis of the matrix. The employee's classification is the union of their x-axis and y-axis position. Employees are classified as "Below", "Meets", or "Exceeds" expectations. No employees were classified in the "N/A" cells.

**TABLE 1**
**List and Description of Proxies**

| Proxy | Description |
|---|---|
| *PRE RATING* | Coded 1, 2, 3 for employees who were rated "below", "meets" "exceeds" expectations prior to the calibration process. |
| *CHANGE DOWN* | Coded 1 if employee received a downward revision, 0 otherwise. |
| *RATING CHANGE* | *POST RATING – PRE RATING* |
| *POST RATING* | Coded 1, 2, 3 for employees who were rated "below", "meets" "exceeds" expectations after the calibration process. |
| *POLITICAL INFLUENCE* [a] $\alpha = 0.76$ | Variable created using PCA on the following items:<br>• I spend a lot of time and effort at work networking with others.<br>• I am good at building relationships with influential people at work.<br>• I have developed a large network of colleagues and associates at work whom I can call on for support when I really need to get things done.<br>• At work, I know a lot of important people and I am well connected.<br>• I spend a lot of time at work developing connections with others.<br>• I am good at using my connections and network to make things happen at work.<br>• Variable coded 1 if supervisor had personally talked to one of the three most influential executive directors over the past couple of years when they wanted to affect the outcome of an important decision, 0 otherwise. |
| *REPUTATION* [a,b] | Variable created for each supervisor by averaging responses to the following question from peer supervisors in their calibration committee:<br>• How important is it to _____ that his/her subordinates are perceived as top performers by others in the company? |
| *LOW PEER SUPPORT* | Coded 1 for those supervisors who are the only one from their functional area in their calibration meeting, contingent upon the calibration meeting having more than two supervisors, 0 otherwise. |
| *AVOID CONFRONTATION* [a,b] $\alpha = 0.87$ | Variable created for each supervisor by averaging responses to the following questions from peer supervisors in their calibration committee. Variable created using PCA on the average responses<br>• Do you think that _____ avoids confrontations with his/her subordinates?<br>• Do you think that _____ finds it hard to criticize his/her subordinates, even if the negative feedback is totally justified?<br>• Do you think that _____ avoids confrontations with other managers? |

For constructs measured using multiple questions, Cronbach's Alpha (α) is provided.

[a] All questions were measured using a 7-point Likert Scale with higher responses indicating higher levels of the variable with the exception of the last item of *POLITICAL INFLUENCE*, which is measured as described.

[b] For the peer evaluation questions, each supervisor answered each question for all other supervisors in the calibration meeting. The name of a supervisor was inserted where the blank is. This process was repeated until the supervisor had answered all peer evaluation questions for all other supervisors in the calibration meeting.

**TABLE 2**
**Descriptive Statistics**

**Panel A: Dependent Variables**

| Variable | Mean | Median | Std. Dev. | Min. | Max. |
|---|---|---|---|---|---|
| *PRE RATING* | 2.13 | 2.00 | 0.54 | 1.00 | 3.00 |
| *CHANGE DOWN* | 0.08 | 0.00 | 0.28 | 0.00 | 1.00 |
| *RATING CHANGE* | -0.06 | 0.00 | 0.34 | -2.00 | 1.00 |
| *POST RATING* | 2.07 | 2.00 | 0.52 | 1.00 | 3.00 |

**Panel B: Independent Variables**

| Variable | Mean | Median | Std. Dev. | Min. | Max. |
|---|---|---|---|---|---|
| *POLITICAL INFLUENCE* | 0.00 | 0.10 | 1.73 | -6.15 | 3.18 |
| *REPUTATION* | 5.63 | 5.75 | 0.98 | 2.00 | 7.00 |
| *LOW PEER SUPPORT* | 0.37 | 0.00 | 0.48 | 0.00 | 1.00 |
| *AVOID CONFRONTATION* | 0.00 | -0.08 | 1.56 | -3.94 | 3.20 |

**Panel C: Control Variables**

| Variable | Mean | Median | Std. Dev. | Min. | Max. |
|---|---|---|---|---|---|
| *CAL SIZE* | 4.44 | 4.00 | 1.82 | 2.00 | 9.00 |
| *EMP JOB* | | | | | |
| *1* | 0.30 | 0.00 | 0.46 | 0.00 | 1.00 |
| *2* | 0.17 | 0.00 | 0.37 | 0.00 | 1.00 |
| *3* | 0.12 | 0.00 | 0.32 | 0.00 | 1.00 |
| *4* | 0.07 | 0.00 | 0.26 | 0.00 | 1.00 |
| *5* | 0.21 | 0.00 | 0.41 | 0.00 | 1.00 |
| *6* | 0.13 | 0.00 | 0.33 | 0.00 | 1.00 |
| *EMP REGION* | | | | | |
| *1* | 0.46 | 0.00 | 0.50 | 0.00 | 1.00 |
| *2* | 0.23 | 0.00 | 0.42 | 0.00 | 1.00 |
| *3* | 0.09 | 0.00 | 0.29 | 0.00 | 1.00 |
| *4* | 0.22 | 0.00 | 0.41 | 0.00 | 1.00 |

Panels A through C provide summary statistics for our dependent and independent variables. Table 1 provides details on the dependent and independent variables. *CAL SIZE* is the number of supervisors present in each meeting. *EMP JOB* and *EMP REGION* are indicator variables for the type of job and region of Brazil in which the employee works.

**TABLE 3**
**Correlation Matrix**

| | *PRE RATING* | *RATING CH.* | *POST RATING* | *POL. INF.* | *REP.* | *LOW PEER* | *AVOID CONF.* |
|---|---|---|---|---|---|---|---|
| *PRE RATING* | 1.00 | | | | | | |
| *RATING CH.* | **-0.36** | 1.00 | | | | | |
| *POST RATING* | **0.84** | **0.21** | 1.00 | | | | |
| *POL. INF.* | 0.01 | 0.03 | 0.03 | 1.00 | | | |
| *REPUTATION* | **0.19** | -0.01 | **0.19** | 0.03 | 1.00 | | |
| *LOW PEER SUP.* | -0.17 | -0.04 | **-0.19** | 0.05 | 0.03 | 1.00 | |
| *AVOID CONF.* | **0.24** | -0.13 | 0.18 | 0.04 | 0.14 | 0.14 | 1.00 |

Bold numbers indicate statistical significance at the 5 percent level or lower (two-tailed). Please see Table 1 for a list of all variable definitions.

**TABLE 4**
**Rating Distribution**

**Panel A: Number (Percent) of Employees Classified as Below, Meets, or Exceeds Expectations**

| | Below | Meets | Exceeds |
|---|---|---|---|
| Required Distribution | 10-20% | 60-80% | 10-20% |
| Pre-Calibration Meeting | 67 (9.1%) | 508 (68.9%) | 162 (22.0%) |
| Post-Calibration Meeting | 78 (10.6%) | 532 (72.2%) | 127 (17.2%) |

**Panel B: Rating Change Detail**

| | Downward Revisions | | | No Change | Upward Revisions | |
|---|---|---|---|---|---|---|
| | From Exceeds to Below | From Exceeds to Meets | From Meets to Below | No Change | From Below to Meets | From Meets to Exceeds |
| Freq. (%) of Changes | 2 (0.3%) | 44 (6.0%) | 16 (2.2%) | 657 (89.1%) | 7 (0.9%) | 11 (1.5%) |

**Panel C: Ratings Across Calibration Meetings**

| CAL | # SUP | PRE | CHANGE | POST | CAL | # SUP | PRE | CHANGE | POST |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2.10 | 0.00 | 2.10 | 15 | 2 | 1.86 | 0.14 | 2.00 |
| 2 | 2 | 1.45 | 0.00 | 1.45 | 16 | 4 | 2.28 | -0.07 | 2.21 |
| 3 | 4 | 2.26 | -0.26 | 2.00 | 17 | 4 | 2.28 | -0.08 | 2.20 |
| 4 | 5 | 2.00 | -0.05 | 1.95 | 18 | 3 | 1.96 | -0.09 | 1.87 |
| 5 | 5 | 1.75 | -0.05 | 1.70 | 19 | 4 | 1.97 | -0.09 | 1.88 |
| 6 | 3 | 1.83 | -0.25 | 1.58 | 20 | 5 | 2.13 | 0.06 | 2.19 |
| 7 | 6 | 2.45 | -0.14 | 2.31 | 21 | 4 | 2.27 | -0.04 | 2.23 |
| 8 | 6 | 2.09 | -0.03 | 2.06 | 22 | 3 | 2.23 | -0.08 | 2.15 |
| 9 | 5 | 2.00 | 0.00 | 2.00 | 23 | 5 | 2.17 | 0.02 | 2.19 |
| 10 | 2 | 2.08 | 0.00 | 2.08 | 24 | 4 | 1.97 | -0.06 | 1.92 |
| 11 | 8 | 2.49 | -0.23 | 2.26 | 25 | 2 | 1.40 | -0.10 | 1.30 |
| 12 | 7 | 2.14 | -0.09 | 2.05 | 26 | 3 | 2.23 | -0.06 | 2.17 |
| 13 | 9 | 2.19 | 0.00 | 2.19 | 27 | 5 | 2.06 | -0.09 | 1.97 |
| 14 | 6 | 2.22 | -0.04 | 2.18 | 28 | 5 | 2.44 | -0.13 | 2.31 |

Panel A provides the number of employees that received "below", "meets", or "exceeds" expectations both pre- and post-calibration meeting. Panel B provides additional detail on changes resulting from the calibration meetings. Panel C provides calibration-level detail about the ratings. CAL is the calibration identifier. # SUP refers to the number of supervisors in each meeting. PRE (CHANGE) [POST] shows the average pre (change in) [post] calibration ratings. Calibration meeting number twenty-five is not included in our final sample because only one of the two supervisors responded to the survey.

**TABLE 5**
**Regression Models for Pre Rating, Rating Change, and Post Rating**

| | **Pre Rating** | **Rating Change** | | | **Post Rating** | |
|---|---|---|---|---|---|---|
| Variable | Ordered Probit | Logit | Logit | OLS | Ordered Probit | Ordered Probit |
| | (1) *PRE RATING* | (2) *CHANGE DOWN* | (3) *CHANGE DOWN* | (4) *RATING CHANGE* | (5) *POST RATING* | (6) *POST RATING* |
| *PRE RATING* | - | - | 2.82*** (7.61) | -0.25*** (-5.41) | - | - |
| *OBJ. PERFORMANCE* | - | - | - | - | - | 0.45*** (3.94) |
| *POLITICAL INFLUENCE* | 0.09** (2.23) | 0.00 (0.04) | -0.06 (-0.53) | 0.01 (1.26) | **0.10*** (3.47)** | **0.08*** (3.46)** |
| *REPUTATION* | 0.21*** (3.25) | -0.01 (-0.02) | -0.18 (-0.79) | 0.01 (0.99) | **0.19*** (2.92)** | **0.12** (2.04)** |
| *LOW PEER SUPPORT* | -0.10 (-0.70) | **0.85** (2.00)** | **1.31*** (3.89)** | **-0.06** (-1.92)** | **-0.24*** (-2.25)** | **-0.22*** (-2.20)** |
| *AVOID CONFRONTATION* | **0.09** (2.12)** | **0.30** (2.07)** | **0.17* (1.27)** | **-0.02** (-1.85)** | 0.02 (0.44) | 0.03 (0.77) |
| *CAL SIZE* | 0.10** (2.14) | 0.01 (0.03) | -0.08 (-0.66) | 0.01 (1.39) | 0.11*** (3.59) | 0.11*** (5.17) |
| Job Indicator | Yes | Yes | Yes | Yes | Yes | Yes |
| Region Indicator | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 737 | 737 | 737 | 737 | 737 | 737 |
| Cal Clusters | 27 | 27 | 27 | 27 | 27 | 27 |
| $R^2$/Psuedo $R^2$ | 0.13 | 0.08 | 0.28 | 0.17 | 0.12 | 0.13 |

*, **, *** indicate significance at the 0.10, 0.05, 0.01 level, respectively. Coefficients and (z-scores) provided. Hypothesized directional relations are in bold and their p-values are one-tailed. All other p-values are two-tailed.

This table provides the output from multiple regression analyses. Table 1 provides details on the dependent and independent variables. *OBJECTIVE PER.* is the objective departmental performance measure. Each employee within a department receives the same base score between one and four contingent on the department's performance. The department supervisor can then adjust the score of each individual employee up or down one point, based on his or her contribution to the department's performance. *CAL SIZE* is the number of supervisors present in each meeting. *EMP JOB* and *EMP REGION* are indicator variables for the type of job and region of Brazil in which the employee works.

**TABLE 6**
**Evidence of Informal Strategic Alliances through Peer Support**

**Panel A: Comparison of ratings for supervisors with and without peer support**

| | *PRE RATING* | *CHANGE DOWN* | *POST RATING* |
|---|---|---|---|
| *Supervisors without peer support* (n = 307 employees; 50 supervisors) | 2.10 (0.03) | 0.11 (0.02) | 2.02 (0.03) |
| *Supervisors with peer support* (n = 430 employees; 63 supervisors) | 2.15 (0.03) | 0.06 (0.01) | 2.10 (0.03) |
| Difference | t = 1.32 p = 0.19 | t = 2.48 p = 0.01 | t = 2.21 p = 0.03 |

**Panel B: Comparison of homogeneous and non-homogeneous calibration committees**

| | *PRE RATING* | *CHANGE DOWN* | *POST RATING* |
|---|---|---|---|
| *Supervisors in homogeneous calibration committees* (n = 297 employees; 39 supervisors) | 2.19 (0.03) | 0.05 (0.01) | 2.16 (0.03) |
| *Supervisors not in homogeneous calibration committees* (n = 440 employees; 74 supervisors) | 2.09 (0.03) | 0.11 (0.01) | 2.00 (0.02) |
| Difference | t = 2.60 p = 0.01 | t = 2.44 p = 0.02 | t = 4.10 p < 0.01 |

Average (Std. Error) are provided. A comparison of the two means is provided in the last row through two-tailed t-tests.

This table provides evidence of the importance of peer support. We argue that those supervisors who have at least one other supervisor from their functional area in their calibration meeting are able to build informal alliances that lead to willingness to support each other during the calibration process, which results in higher employee performance ratings. For Panel A, within each calibration meeting, we code supervisors that do have (do not have) at least one other supervisor in their functional area as *Supervisors with (without) peer support*. We then compare *PRE RATING*, *CHANGE DOWN*, and *POST RATING* across the two subsamples. For Panel B, we code each calibration committees based on whether it is homogenous or not. In particular, we code those calibration committees that only have supervisors from a single functional area as *Homogenous calibration committees*.

**TABLE 7**
**Mixed-Effect Regression Models for Pre Rating, Rating Change, and Post Rating**

| | **Pre Rating** | **Rating Change** | | | **Post Rating** | |
|---|---|---|---|---|---|---|
| | Multilevel mixed-effects ordered probit regression | Multilevel mixed-effects logistic regression | Multilevel mixed-effects logistic regression | Multilevel mixed-effects linear regression | Multilevel mixed-effects ordered probit regression | Multilevel mixed-effects ordered probit regression |
| | (1) *PRE RATING* | (2) *CHANGE DOWN* | (3) *CHANGE DOWN* | (4) *RATING CHANGE* | (5) *POST RATING* | (6) *POST RATING* |
| *PRE RATING* | - | - | 3.01*** (7.27) | -0.24*** (-10.05) | | - |
| *OBJ. PERFORMANCE* | | | | | | 0.49*** (3.89) |
| *POLITICAL INFLUENCE* | 0.09* (1.80) | -0.03 (0.21) | -0.08 (-0.69) | 0.01 (1.10) | **0.10*** (2.80)** | **0.08*** (2.24)** |
| *REPUTATION* | 0.28*** (3.18) | -0.08 (-0.32) | -0.35 (-1.37) | 0.02 (1.27) | **0.27*** (3.64)** | **0.15*** (2.25)** |
| *LOW PEER SUPPORT* | -0.21 (-1.13) | **0.78** (1.63)** | **1.36*** (2.87)** | **-0.06** (-1.60)** | **-0.25** (-1.71)** | **-0.25** (-1.90)** |
| *AVOID CONFRONTATION* | **0.11** (2.02)** | **0.36*** (2.24)** | **0.21* (1.40)** | **-0.02** (-1.80)** | 0.01 (0.27) | 0.04 (1.04) |
| *CAL SIZE* | 0.10* (1.89) | -0.11 (-0.76) | -0.18 (-1.19) | 0.02* (1.65) | 0.12** (2.36) | 0.12*** (2.92) |
| Job Indicator | Yes | Yes | Yes | Yes | Yes | Yes |
| Region Indicator | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 737 | 737 | 737 | 737 | 737 | 737 |
| Wald $\chi^2$ | 98.78 | 17.38 | 62.36 | 119.03 | 93.70 | 107.44 |
| Prob > $\chi^2$ | < 0.01 | 0.18 | < 0.01 | < 0.01 | < 0.01 | < 0.01 |

*, **, *** indicate significance at the 0.10, 0.05, 0.01 level, respectively. Coefficients and (z-scores) provided. Hypothesized directional relations are in bold and their p-values are one-tailed. All other p-values are two-tailed.

This table provides the output from multiple regression analyses. Please see Table 1 for details on the dependent and independent variables. *CAL SIZE* is the number of supervisors present in each meeting. *EMP JOB* and *EMP REGION* are indicator variables for the type of job and region of Brazil in which the employee works.